基於混合式長短期記憶網路架構之空氣污染預測

（一） 摘要

隨著近年來國人健康意識抬頭，空氣汙染已經成為眾人關注的議題。除了氣象預報以外，空氣品質指標（AQI）也成為日常生活中不可或缺的項目，其中考量多種空氣汙染因子，如：臭氧(O_3)、細懸浮微粒($PM_{2.5}$)、懸浮微(PM_{10})、一氧化碳(CO)等。其中「細懸浮微粒($PM_{2.5}$)」對人體的影響尤為嚴重。目前為止，在空氣汙染預測領域的相關研究不在少數，但鮮少有研究將氣象相關資訊與空氣汙染因子的移動一起探討。空氣汙染濃度可能受周遭地區降雨、風向、風速等影響，若能將氣象因子加入模型中，預期能提高預測的準確度。本計畫分為兩階段。第一階段使用長短期記憶模型（LSTM）與注意力機制（attention mechanism），透過考慮風向與降雨量的空間變化對鄰近地區 $PM_{2.5}$濃度的影響，藉此預測目標地點的$PM_{2.5}$濃度。基於附近的氣象與污染濃度資料，通過 LSTM 神經網路得到初步的 $PM_{2.5}$預測結果。第二階段使用 XGBoost (eXtreme Gradient Boosting, 極限梯度提升)將第一階段的初步預測結果與天氣預測結果相結合，以得到對於 $PM_{2.5}$濃度的最終預測結果。

關鍵字: 空氣汙染預測、深度學習。

（二） 研究動機與研究問題

自從工業發展以來，空氣汙染問題愈發嚴重，已成為全球關注的問題。根據世界衛生組織的數據，全球十大死亡原因中，有三項為呼吸道疾病[1]，分別為:第三名-慢性阻塞性肺炎、第四名-下呼吸道感染、第六名-氣管、支氣管、肺癌。如圖 1 所示。在所有汙染物中，$PM_{2.5}$是危害健康最大的因素，可以透過鼻子、喉部進入氣管及肺泡，可能導致嚴重的呼吸道疾病[2]，連帶影響國人就醫比率[3]，增加醫療負擔[4]。因此，監控與控制空氣汙染物已成為重要的課題。我國環保署（EPA）設置 84 個中央空氣品質測站，每小時更新各汙染物的濃度，民眾可上網查看即時的空氣汙染資訊，以便決定是否外出。因此，本計畫研究動機為考量 $PM_{2.5}$的威脅日漸嚴重，需要一個可信任且精準的空汙預測系統，以便提前發出預警。

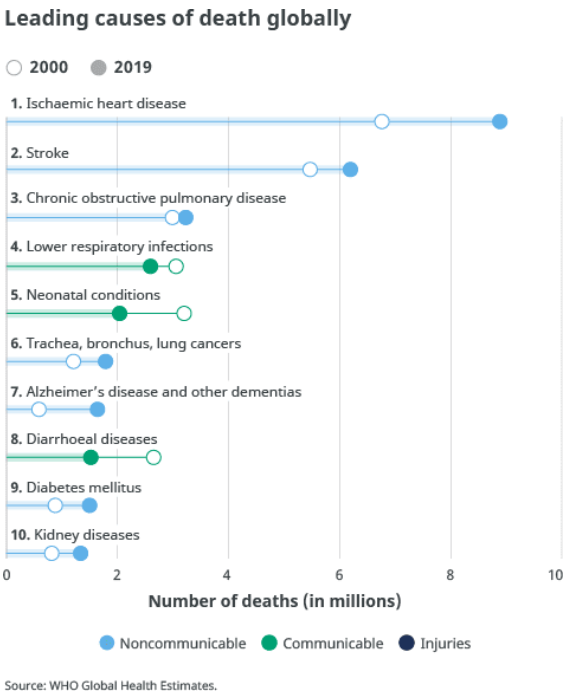

圖 1 - WHO 統計 2019 年全球十大死因

空氣汙染的分布會受天氣與人類活動影響，而在時間與空間有連續性的變化[5]，而天氣變化對空氣汙染的影響尤其之大[6]，可能受氣溫、濕度、降雨、風速、風向等影響。空氣汙染預測是一個標準的時間序列問題。在機器學習與深度學習領域中，常應用在解決時間序列問題的模型有遞迴神經網路（RNN）、長短期記憶模型（LSTM）等。以往對於空氣汙染預測往往只基於空氣污染的歷史資料[7][8]，鮮少將氣象因子一同加入考慮，因此沒有辦法很有效預測未來天氣變化對空氣污染程度的影響。目前天氣預報的可信度已大幅提升，如果能將天氣預報良好運用在空氣汙染的預測中，則可以在一定程度上提高預測的準確性。因此，本計畫研究問題為應用 LSTM 模型考慮氣象因素得出第一階段 $PM_{2.5}$ 預測結果，再結合氣象預報資料，並應用XGBoost模型得出進一步預測結果。

（三） 文獻回顧與探討

3.1 空汙預測與空氣汙染因子

影響空氣汙染濃度變化的因子包含境外汙染、天氣等因素。Borge et al.(2019) 研究發現，溫度和降雨對空氣汙染源濃度有一定程度的影響。Roger et al.(2005)在研究中也表示，$PM_{2.5}$ 濃度有季節性的差異，其中冬季濃度最高，夏季最低。機器學習已經廣泛運用在空氣汙染的預測，林等(2018)[10]使用三種不同的機器學習演算法分別訓練並建立模型，其中以支援向量回歸（Support Vector Regression，簡稱 SVR）的結果最為準確。SVR 是迴歸分析的一種，不同於線性回歸，SVR 是找出一個最佳的超平面來預測數值。楊等（2018）使用 XGBoost，先將空氣汙染進行分類，再對空氣汙染的類別進行預測。XGBoost 是基於 Gradient Boosted Decision Tree (GBDT) 改良與延伸，被應用於解決監督式學習的問題。可透過調整學習率(learning rate)來解決過度擬合(overfitting)的問題，且提供並行樹的梯度提升，因此可以快速解決數據科學的問題。

3.2 時間序列問題

可藉由分析時間序列的趨勢、循環等，藉此預測未來時間的數值。早期透過遞歸神經網路（RNN）來研究[12]，LSTM 即是一種遞迴神經網路，擅長處理和預測時間序列問題，廣泛被運用在空汙預測問題中[13]，但卻很少有研究討論由風造成空汙影響的問題。RNN 模型如圖 2 所示。

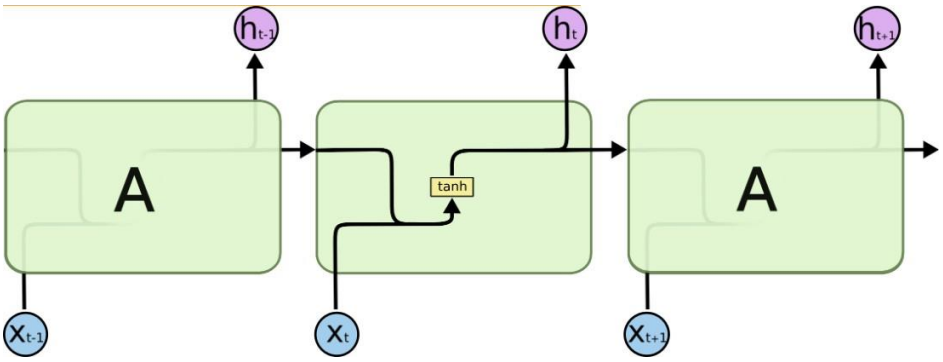

圖 2 - RNN 模型示意圖

3.3 長短期記憶模型 (LSTM)[14][15]

與傳統 RNN 不同，LSTM 內部有四層結構。如圖 3 所示。第一層為 forget gate，用來決定哪些資訊要從細胞狀態中被遺忘。第二部分為 input gate，決定哪些新資訊要被更新到細胞中，然後還有一層 tanh 函數將細胞更新。最後是 output gate，透過 tanh 控制該層細胞有多少資訊繼續做為下一次的輸入。LSTM控制前一個細胞的資訊傳遞，能有效解決傳統RNN所造成之梯度消失問題(Vanishing gradient problem)。因為空氣汙染濃度會隨時間有連續性的變化，因此預期LSTM在此問題上能有良好的表現。

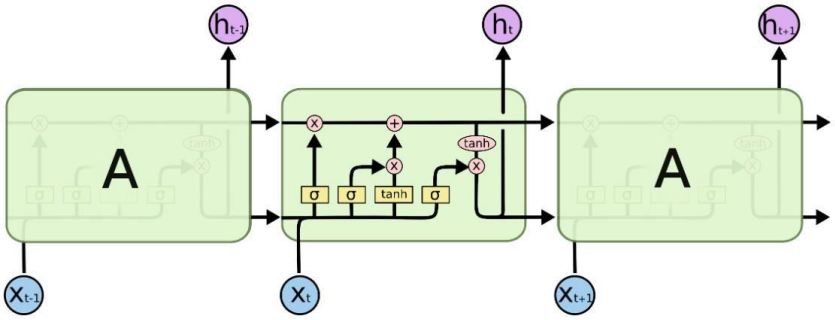

圖 3 - LSTM 模型示意圖

（四） 研究方法與步驟

4.1 研究模型概述

本研究模型由四個部分組成：資料前處理、風敏長短期記憶模型、降雨長短期記憶模型、集成學習預測。如圖 4 所示。

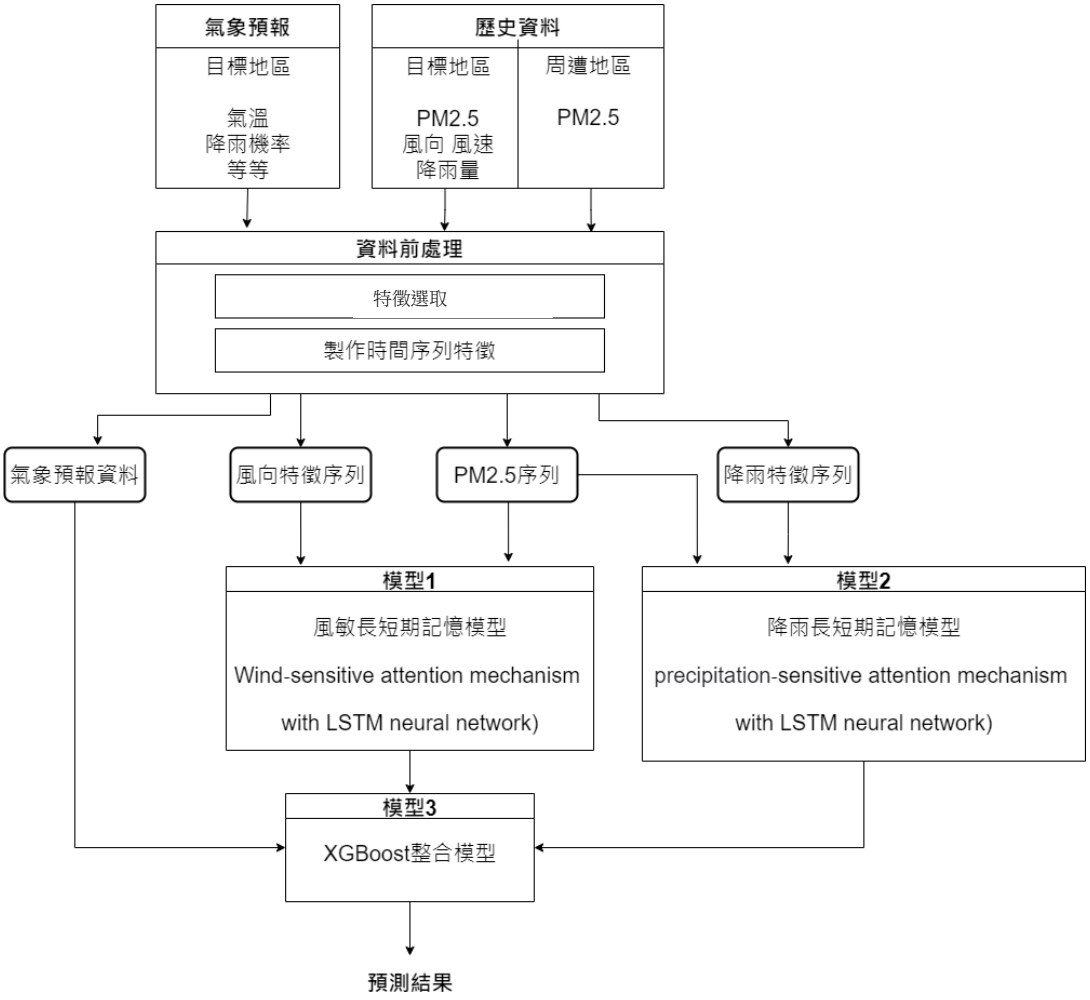

圖 4 – 研究模型架構圖

4.2 資料前處理(data preprocessing)

為了利用天氣特徵（尤其是風）協助 $PM_{2.5}$ 的時間序列預測，如何表示風的特徵及組合數據顯得格外重要。

4.2.1 對風向特徵進行獨熱編碼

EPA 的資料中，風向是以 0° 到 360°表示，但略大於 0°與略小於 360°的風向會被判定為相反方向，因此先將風向分類為「北」、「東北」、「東」…「西北」八個方向，接著使用獨熱編碼(One-hot encoding) 表示之。如式(1)-(3)：

$$f(x,c)=\begin{cases}1, x=c\\ 0, x\neq c\end{cases} \tag{1}$$

$$f_{onehot}(x,S)=\langle f(x,c_1),f(x,c_2),\cdots,f(x,c_S)\rangle \tag{2}$$

$$\vec{wd}=f_{onehot}(x_{wd},S) \tag{3}$$

4.2.2 製作時間序列特徵

考慮當前時間 ct 的過去 m 個時間段與未來 n 個時間段，產生與風相關的時間序列。如式(4)。

$$\overrightarrow{Wind}_{ct}=\langle\overrightarrow{Wind}_{ct-m+1},\overrightarrow{Wind}_{ct-m+2},\cdots,\overrightarrow{Wind}_{ct},\cdots,\overrightarrow{Wind}_{ct+n}\rangle \tag{4}$$

降雨量考慮過去時間段的資料，除了時間以外，同時也要將周遭區域資料列入考慮。如式(5)。

$$\overrightarrow{Prec}_{ct}=\begin{bmatrix}Prec_{1,ct-k+1} & \cdots & Prec_{L,ct-k+1}\\ \vdots & \ddots & \vdots\\ Prec_{1,ct} & \cdots & Prec_{L,ct}\end{bmatrix} \tag{5}$$

$PM_{2.5}$ 濃度只有考慮過去時間段的資料，除了時間以外，同時也要將周遭地點列入考慮。如式(6)。

$$\overrightarrow{PM2.5}_{ct}=\begin{bmatrix}PM2.5_{1,ct-k+1} & \cdots & PM2.5_{L,ct-k+1}\\ \vdots & \ddots & \vdots\\ PM2.5_{1,ct} & \cdots & PM2.5_{L,ct}\end{bmatrix} \tag{6}$$

4.3 風敏長短期記憶模型(Wind-sensitive attention mechanism with LSTM neural network)

本模型使用注意力機制（attention mechanism）的核心概念，關注風向對 $PM_{2.5}$ 濃度的影響，並將得到的特徵權重用於 LSTM 模型的預測中。如圖 5 所示。

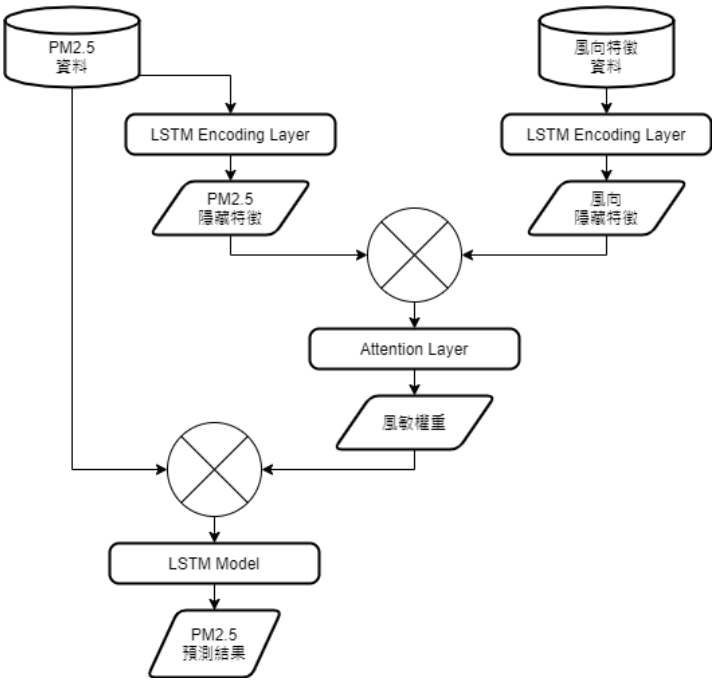

圖 5 - 風敏長短期記憶模型流程圖

4.4 降雨長短期記憶模型(Precipitation attention mechanism with LSTM neural network)

本模型使用注意力機制（attention mechanism）的核心概念，關注周遭地區降雨量對 $PM_{2.5}$ 濃度的影響，並將得到的特徵權重用於 LSTM 模型的預測中。如圖 6 所示。

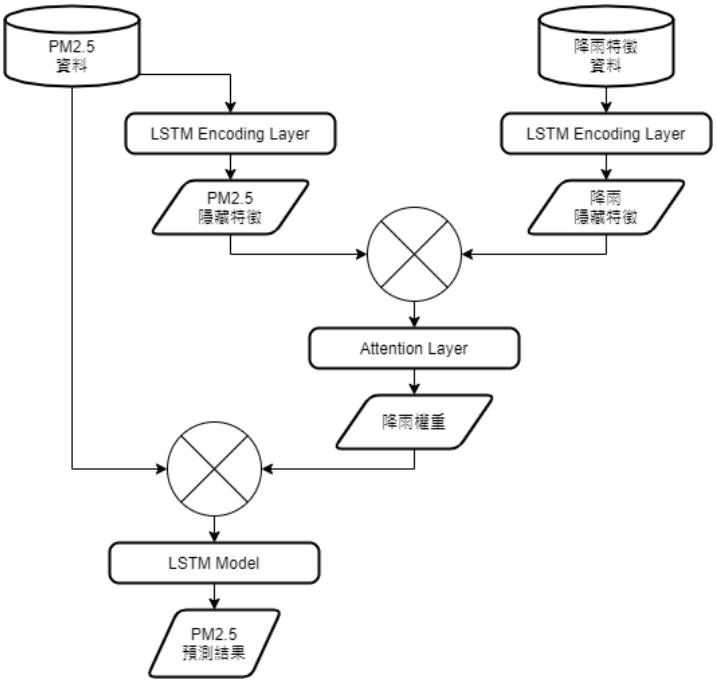

圖 6 - 降雨長短期記憶模型流程圖

4.5 運用 XGBoost 進行第二階段預測

在這個階段，將氣象預報資料與前一階段獲得的 $PM_{2.5}$ 濃度預測應用 XGBoost模型，做第二階段的 $PM_{2.5}$ 濃度預測。

4.5.1 取得氣象預報特徵

氣象局所提供的氣象預報資料，包含風速、溫度、濕度、降雨機率等。用這些資料產生天氣特徵，如式(7)。

$$\overrightarrow{\text{weather}}_t = \langle x_{Temp,t}, x_{RH,t}, x_{Rain,t}, x_{WS,t} \rangle \quad (7)$$

4.5.2 XGBoost 模型

XGBoost 是一種提升演算法，能將一系列較弱的學習演算法整合得到一個強學習演算法的結果。經過第一階段的預測後，將數個風敏長短期記憶模型與降雨長短期記憶模型加上氣象預報特徵資料結合，以得出第二階段預測結果。如式(8)。

$$\hat{y}_t = \sum_{tree=1}^{T} f_{tree}(\overrightarrow{y_{t\cdot}}, \overrightarrow{\text{weather}}_t) \quad (8)$$

（五） 預期結果

本計畫中使用基於注意力機制與 LSTM 神經網路的空氣汙染 $PM_{2.5}$ 濃度預測模型，可預測 6 至 24 小時內的 $PM_{2.5}$ 濃度。不同於以往完全依賴歷史資料的預測方式，加入天氣預測資料後，可更準確預測未來空氣汙染程度，同時也受惠於未來氣象預報技術的持續進步。期望未來能幫助政府與國民加強對空氣汙染的警覺與自我保護的能力。本計畫著重於 $PM_{2.5}$ 濃度與風向風速的相互作用關係，未來可再加入更多參數計算(如:地形)，改善預測的準確度。

（六） 參考文獻

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Xing Y.F., Xu Y.H., Shi M.H., Lian Y.X. (2016)The impact of PM2.5 on the human respiratory system.

[3] 黃鈺茹（2015）。以追蹤資料分析臺灣空氣污染與呼吸道及心血管疾病之相關性。國立臺北科技大學環境工程與管理研究所碩士論文，台北市。取自https://hdl.handle.net/11296/zeepxz

[4] 余昌航（2009）。空氣污染物及氣象因子與呼吸道疾病門診量之相關性研究－以台北市為例。中華技術學院土木防災工程研究所在職專班碩士論文，台北市。 取自 https://hdl.handle.net/11296/w4vqus

[5] Huang, P., Zhang, J., Tang, Y., & Liu, L. (2015). Spatial and temporal distribution of PM2. 5

pollution in Xi'an City, China. International Journal of Environmental Research and Public Health, **12**(6), 6608 - 6625.

[6]Borge R., Requia W.J., Yagüe C., Jhun I., Koutrakis P.（2019） Impact of weather changes on air quality and related mortality in Spain over a 25 year period [1993 - 2017]

[7] 黃偉勝（2017）。以時間序列資料之模糊預測-以台中市空氣污染公開資料為例。東海大學資訊工程學系碩士論文，台中市。 取自 https://hdl.handle.net/11296/n9cutn

[8] 楊宏宇（1993）。臺灣地區空氣品質與天氣類型分類相關性研究。文化大學地學研究所博士論文，台北市。 取自 https://hdl.handle.net/11296/mn8qup

[9] Roger D. Peng, Francesca Dominici, Roberto Pastor-Barriuso, Scott L. Zeger, Jonathan M. Samet(2005), Seasonal Analyses of Air Pollution and Mortality in 100 US Cities, American Journal of Epidemiology,Volume 161,Issue 6,15 March 2005,Pages 585-594, https://doi.org/10.1093/aje/kwi075

[10] 林冠名（2018）。在大數據平台使用機器學習方法預測空氣汙染。國立臺北大學資訊工程學系碩士論文，新北市。 取自 https://hdl.handle.net/11296/dvxvgw

[11] 楊軒（2018）。基於時間序列、迴歸和正規化的快速預測 PM2.5 方法。國立臺灣師範大學資訊工程學系碩士論文，臺北市。

[12] X. ZhaoR. ZhangJ.-L. WuP.-C. Chang (2018)A deep recurrent neural network for air quality classification

[13] Xayasouk, T., Lee, H., & Lee, G. (2020). Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. Sustainability, 12(6), 2570. doi:10.3390/su12062570

[14] Hochreiter, S.,& Schmidhuber,J.(1997).Long short-term memory.Neural Computation,**9**(8), 1735 - 1780.

[15] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

（七） 需要指導教授指導內容

1. 研究方法指導
2. 演算法建模訓練
3. 實驗設計與結果驗證分析